MEMORANDUM
RM-5387-PR
NOVEMBER 1967

DIGITAL COMPUTER SIMULATION: STATISTICAL CONSIDERATIONS

George S. Fishman and Philip J. Kiviat

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

MEMORANDUM

RM-5387-PR

NOVEMBER 1967

DIGITAL COMPUTER SIMULATION: STATISTICAL CONSIDERATIONS

George S. Fishman and Philip J. Kiviat

This research is supported by the United States Air Force under Project RAND—Contract No. F44620-67-C-0015 monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force.

The RAND Corporation

1700 MAIN ST • SANTA MONICA • CALIFORNIA • 90406

PREFACE

This is one of a series of RAND Memoranda on digital computer simulation. Preceding work on this subject has been described in G. S. Fishman, Digital Computer Simulation: The Allocation of Computer Time in Comparing Experiments, The RAND Corporation, RM-5288-1-PR, October 1967, and P. J. Kiviat, Digital Computer Simulation: Modeling Concepts, The RAND Corporation, RM-5378-PR, September 1967. The purpose of this Memorandum is to describe a number of statistical problems that materialize during computer simulation experiments. The Memorandum gives references (when they exist) that will assist an experimenter in resolving these problems.

SUMMARY

This Memorandum describes a number of statistical problems that arise in computer simulation experiments. Failure to resolve these problems adequately can significantly degrade the value of experimental results. References are given that should assist an experimenter in handling them.

The Memorandum describes three principal problem areas: verification, validation, and problem analysis. Verification insures that a simulation model containing a mathematical structure and a data base behaves as an experimenter intends. The complexity of models often makes it difficult to determine whether their basic operating assumptions are satisfied.

Validation tests the agreement between the behavior of a simulation model and the observed behavior of a real system. This requires empirical data. If a behavioral equivalence can be established between a simulation model and a real system, we may regard the behavior of the model and the system as being consistent. Since a simulation model is often exercised with modifications that do not currently exist in a real system, it is important that a benchmark of consistency be established whenever possible to provide confidence for extrapolations.

Problem analysis embraces a host of statistical problems relating to the collection, reduction, and presentation of data generated by computer simulation. The choice of sampling interval, the use of variance reduction techniques, and the estimation of reliability are problems common to all simulation experiments containing random phenomena. These and similar problems are considered and references are given to discussions and solutions.

CONTENTS

I. INTRODUCTION

Many system simulation experiments are driven by input processes containing elements of random behavior. In such simulations, statistical reliability must be considered if experimental results are to be interpreted properly. Statistical considerations also enter into the evaluation of simulation model designs. This Memorandum describes these considerations, identifying how and where they become important during the planning, performance, and analysis of simulation experiments. The description can be viewed as tracing the elements of a typical experiment from inception through analysis, defining statistical problems and relating them to the formal body of statistical theory.

The problems described are inherent in all stochastic system simulation models. An experimental design's ability to reveal useful insights into a system depends to a great extent on how well these problems are solved. Failure to deal with them may cause errors in interpreting observed associations between system input and output. One common error is the underestimation of the reliability of system response measurements, caused by failure to account for autocorrelation in system response time series generated by a simulation model. Another frequent source of error is the assumption that random numbers generated within a simulation model are independent, when in fact the method of random number generation employed induces unwanted correlation.

Our aim is to promote awareness of problems, not to solve them. The study offers no general solutions, but provides references germane to the statistical problems described. Some references describe particular solutions; others offer methods of analysis.

To understand the role of statistics in system simulation experiments, a knowledge of how these experiments developed is helpful. System simulation may be regarded as an extension of Monte Carlo methods. These methods, which concern experiments with random numbers, began their systematic development during World War II when they were applied to problems related to the atomic bomb. The work involved direct simulation of probabilistic problems concerned with random neutron diffusion

in fissionable material [11]. Shortly thereafter, it was proposed that Monte Carlo methods be applied to solve certain integral equations, occurring in physics, that were not amenable to analytical solution. Stochastic processes often existed whose parameters satisfied these equations. One could estimate these parameters (and hence the solution to the equations) by performing Monte Carlo experiments on the stochastic processes.

The reliability of parameter estimates was the dominant statistical problem in these Monte Carlo experiments. Since the estimates were generally the sum of independent, identically distributed random variables, their reliability was inversely proportional to $n^{\frac{1}{2}}$--a 10-percent improvement in reliability required a 100-fold increase in sample size. For many problems, random sampling was prohibitively expensive even with digital computers. The crucial statistical problem was finding ways of reducing the variance of an estimator for a given sample size. A number of these variance reduction methods are described in [23]. A particularly useful variance reduction technique known as the method of antithetic variates is described in Hammersley and Handscomb [11].

The concept of system simulation became a reality in the early 1950's, when there was a shift in emphasis from looking at parts of a problem to examining the simultaneous interactions of all parts. This shift was at least partially due to the fact that system simulation experiments had become feasible on digital computers, which were undergoing order-of-magnitude advances in speed. Simulation made it possible to carry out fully integrated system analyses which were generally far too complex to be carried out analytically. This was especially true for studies of the interactions among parts of a system.

In the past decade, the ability to model complex systems has greatly improved. Specialized computer simulation languages such as GPSS, SIMSCRIPT and SIMULA offer convenient formats for describing system problems. Along with the improvements, however, have come a number of statistical problems, few of which have been satisfactorily solved. In fact, some of them have not even been recognized yet as serious problems.

Verification, validation, and problem analysis are tasks demanding careful statistical consideration. Verification determines whether a model with a particular mathematical structure and data base actually behaves as an experimenter assumes it does. Validation tests whether a simulation model reasonably approximates a real system. Problem analysis seeks to insure the proper execution of the simulation and proper handling of its results; consequently it deals with a host of matters: the concise display of solutions, efficient allocation of computer time, proper design of tests of comparison, and correct estimates of sample sizes needed for specified levels of accuracy.

In other words, verification and validation insure that a simulation model is properly designed; only after a model has been verified and validated can an experimenter justifiably use a model to probe system behavior. Problem analysis mainly deals with the results of experimental probing.

Of the remaining sections of the Memorandum, Sec. II provides some necessary definitions and motivation, Secs. III and IV discuss problems associated with the design and proof-testing of a simulation model, and Sec. V considers problems associated with the use of simulation models. The format of the last three sections is: presentation of problems, brief discussion of advised solutions, references to relevant literature.

II. SIMULATION MODELS

The concepts discussed from here on can best be understood in the context of a typical simulation model. This section defines a number of terms used in succeeding sections, examines a typical model to show these terms in their proper context, and indicates some problem areas connected with model structure and data systems that should concern every model-builder.

Every simulation model comprises two systems -- a data system and a logical system. Both present a model-builder with problems; both contribute equally to the validity of a final simulation model.

When we first look at a simulation model we see its logical structure--the way in which a system's operations have been analyzed and factored into discrete units, and these units combined so that the model can be made to reproduce the system's behavior. When we look at a model more deeply, we see that it contains sequences of data comparisons and logical tests. These tests cause a model to take different actions depending on numerical values that are either input from the world outside its boundaries or computed within. The model's behavior is conditioned by these data values, and its results are sensitive to data representations and methods of data generation.

Consider the simple one-machine shop with a waiting line, shown in Fig. 1. Items arrive at the machine for processing; the arrow coming from the left shows the jobs arriving with average arrival rate λ. If the machine is free when a job arrives it immediately begins service, which is performed at an average service rate μ. A job that arrives when the machine is engaged waits in line until it can be processed. The waiting line is pictured as a box; in a real system it might be a tote box or a pile of partially completed parts. When a job is completed it leaves the service facility (arrow going to the right), freeing the machine for another job. If jobs are

waiting in the line (queue), one is selected for service according to a queue discipline and the machine is engaged again. If no jobs are waiting, the machine remains idle until the next job arrival.

Fig. 1 -- Simple machine shop model

Systems such as this, in which jobs arrive, possibly wait in queues, and are serviced are called queueing systems. Almost all simulation models have queueing systems imbedded in them.

Simulating a system like the one described requires the definition of events that take place during its operation. Events occur at points in time when system activities begin and end; if an activity has no duration, e.g., a decision made at an instant in time, it only has one related event.

A gross representation of the logical structure of a queueing system is shown in Fig. 2. The activities pictured are jobs arriving and jobs being serviced. Jobs arrive at the shop at random times. Let the first N jobs that arrive be denoted $j_1, j_2, \ldots, j_N$, and their arrival times be denoted $t_1, t_2, \ldots, t_N$. Then the times between job arrivals are: $d_1 = (t_1 - t_0)$, $d_2 = (t_2 - t_1)$, $\ldots$, $d_N = (t_N - t_{N-1})$. Inputs to the queueing system are simulated by generating job arrivals at the service facility; interarrival times rather than arrival times are usually used. When a job arrives, the time when the next job will arrive is computed by random sampling from an interarrival time distribution. Two data problems associated with this simulation are determining the correct statistical sampling distribution and generating

random samples from it. Section IV discusses some problems concerned with selecting a sampling distribution. Methods for generating random samples from various statistical distributions can be found in [4] and [24].

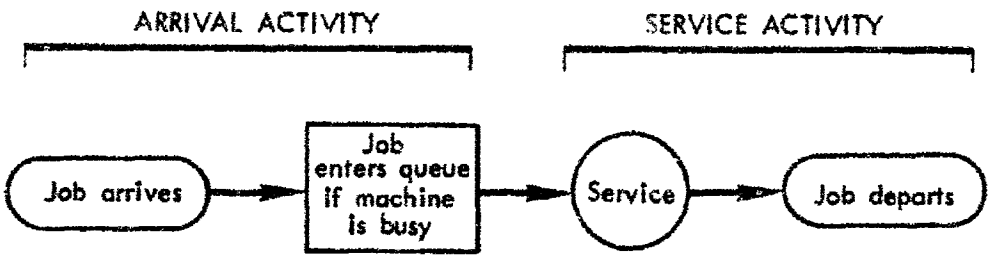

Fig. 2 -- Basic queueing model

A sequence of job arrival times constitutes a sample from a simulation input process. Each arrival generates an interarrival time for the next job and a service time for itself. Figure 3 illustrates the arrival event in some detail showing the sequence of simulation activities: the generation of an interarrival time and a service time, and placement of a new arrival in process or in queue.*

When a job arrives it is placed in service if the server is free; otherwise, it is placed in queue. Call the service times for the N jobs that enter the shop $s_1, s_2, \ldots, s_N$. The sequence of service times also constitutes a simulation input process, as random samples are drawn from some service time distribution whenever a job is processed. For each job that passes through the shop, two (random) quantities must be determined -- d_i and s_i. The quantity d_i determines

*The notation used in Fig. 3 is taken from P. J. Kiviat, Digital Computer Simulation: Modeling Concepts, RM-5378-PR, September 1967.

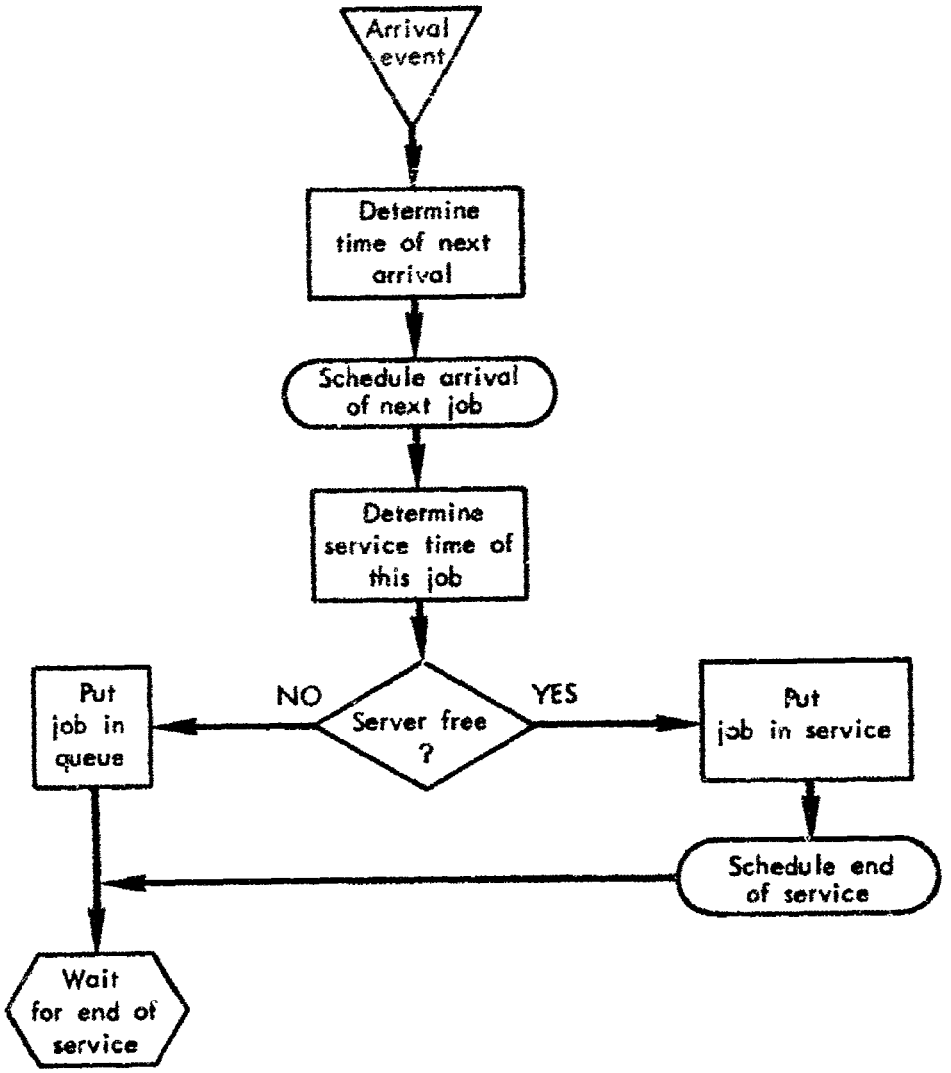

Fig. 3 -- Arrival event

the time when a job enters the shop, s_i determines the time it spends in process. In this model, both of the quantities d_i and s_i are generated when a job enters the shop; in a slightly modified version of the model the service time might not be generated until the job is actually put into service. Regardless of the time (or place in the model) when these values are generated, they "belong" to the job and determine its experience in the system.

All simulation models are driven by some basic force, generally the arrival of a task, job, or request of some sort in the simulated system. Each job's progress through the system is determined by two sets of factors: its characteristics, and pressures exerted by the system. Job characteristics can be few or many; in our simple model there are two, an arrival time and a service time. These characteristics can be generated at one time or at different stages in a job's life as it passes through a simulated system. Regardless of where they are generated, they belong to a job and contribute to its simulated behavior.

A job with n characteristics can be described by a list of these characteristics which we call an n-tuple. A job in our queueing model is characterized by a 2-tuple (d_i, s_i). A typical problem encountered when constructing a simulation model is the generation of job characteristics; an important problem encountered while checking out a simulation model is the examination of a sequence of generated job characterizations, as we call these n-tuples.

As Fig. 3 shows, a job does not necessarily have to pass directly through the shop; it can wait in line while other jobs are being processed. If T_i denotes the time that job i leaves the shop, then $w_i = T_i - t_i - s_i$ is the time it spends waiting for service. The sequences $T_1, T_2, \ldots, T_N$ and $w_1, w_2, \ldots, w_N$ are simulation output processes, sequences of variables whose values are determined by the activities that take place within the simulation model. If the model is designed so that certain jobs have priority over others, then low-priority jobs will have long waiting times; if it is designed with a service facility that shuts down periodically for repairs and rest periods, then the sequence of jobs that exit from the shop will reflect this.

A simulation model is designed to generate output processes that can be studied to observe a system's behavior as its data and/or logical structure are changed. Data influence a model through the selection of statistical sampling distributions, random sampling procedures, and activity levels. The rate at which jobs arrive and are serviced, λ and μ respectively in Fig. 1, are activity levels

that specify the intensity of system operations. Figures 3 and 4 illustrate some influences that model structure exerts on a simulation study.

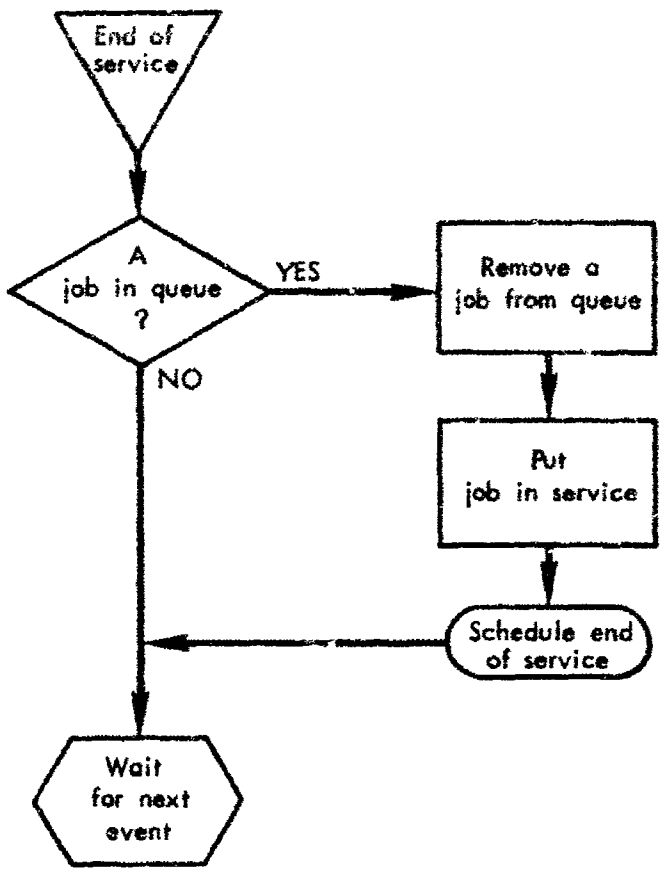

Fig. 4 -- End of service event

The operating rules used to select a job from a waiting line clearly are part of the model structure and influence system behavior. A complex model generally contains many different kinds of operating rules: decision mechanisms, search and choice procedures, and scheduling heuristics are some that are found most frequently. We have chosen a queue discipline to illustrate the effect of an operating rule in a model. A rule under which jobs that have short processing times are selected first will produce a sequence of T_i's close to one

another followed by sequences with greater values. The character of the output process will be different under this rule from the output under a rule that selects jobs in another way.

A simulation model must therefore be examined in two ways. Its data must be examined, both with respect to the particular representations chosen and the way the model selects samples in its simulation process; and its structure must be examined to see that mechanisms have been chosen that produce correct system response. Both data and structure are important, both pose statistical problems in analysis and evaluation. Section III treats in detail the problems outlined in the above example.

III. VERIFICATION

DATA VERIFICATION

Inputs in most simulation experiments consist of jobs of some sort, each characterized by a sequence of random variables. In the simple queueing model each job is characterized by an interarrival time and a service time. Each job affects the system to an extent determined in part by the values of its corresponding 2-tuple. In general, simulation experiments measure the response of a system to different sequences of input n-tuples.

In most system simulation models the elements of job n-tuples are independent random variables and sequences of n-tuples are independent multivariate random variables. The n-tuple elements are transformations of pseudorandom numbers drawn from a uniform distribution on the unit interval. To each n-tuple characterizing a job, there corresponds an n-tuple of uniformly distributed random variables. If the model design is proper, the elements of this latter n-tuple should be independent and uniformly distributed, and so should be the sequences of these n-tuples.

Absence of independence in generated samples implies that the assumptions of the model do not hold. Verifying independence assumptions is the first statistical problem arising in system simulation experiments. Since the tests of independence in no way relate to proposed system structure, one may check the pseudorandom number generator quite separately from other considerations.

The most important hypothesis to test is that the pseudorandom number generator creates sequences of independent random variables. Suppose we collect m pseudorandom numbers. If we divide the unit interval into k class intervals and let x_i be the number of observations in interval i, then for sufficiently large m we may regard the statistic

$$X^2 = \frac{k}{m} \sum_{i=1}^{k} x_i^2 - m$$

as being χ^2 distributed with (k - 1) degrees of freedom.

Mann and Wald [17], who have studied the problem of choosing k according to some "best criterion," suggest

$$k = 4[2(m - 1)^2/c^2]^{1/5},$$

where

$$(2\pi)^{-1/2} \int_c^{\infty} e^{-x^2/2}\, dx = \alpha.$$

Cochran [1] describes the sense in which this choice of k is best. For our purposes, the Mann and Wald criterion seems reasonable. If X^2 exceeds $\chi^2_{k-1,\,\alpha}$, α being the confidence level, we reject the hypothesis. This test or an equivalent one has been performed on most pseudorandom number generators and, therefore, our mentioning it is principally for completeness.

The χ^2 test also applies in testing the independence of n-tuples, but instead of working with the unit interval we divide the n-dimensional unit surface into k n-cubes of equal volume and define x_i as the number of n-tuples in the i^{th} n-cube. MacLaren and Marsaglia [16] apply this test to the output of several pseudorandom number generators for pairs and triples. Their results show a number of standard generators to be suspect.

The χ^2 test concerns questions of randomness and makes no use of the way in which a particular method generates random numbers. Coveyou and MacPherson [5], who offer a unified theory of the statistical behavior of n-tuples of pseudorandom generators, conclude that currently there is no better method of generating n-tuples than the simple multiplicative congruence method, $r_{i+1} = r_i U (\mathrm{mod}\ 2^p)$, with a carefully

chosen multiplier, U. They describe how to choose the multiplier, and discuss the effects of computer word length on generated sequences.

A departure from the independence assumption can significantly affect experimental results. The following example illustrates this point. Let x and y be pseudorandom numbers that are suitably transformed; g(x) is used as an interarrival time and n(y) as a service time. Figure 5 shows the square over which the pair x, y are uniformly distributed.

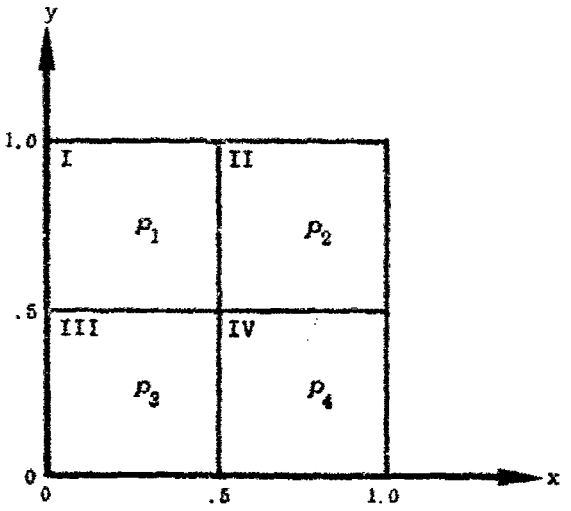

Fig. 5

Let p_i be the probability of x, y being in the i^{th} square. If pairs are independent, then

$$p_i = 1/4 \qquad i = 1, 2, \ldots, 4.$$

Suppose, however, that p_1 is greater than p_2, p_3 and p_4. If interarrival and service times are increasing functions of x and y, respectively, then we would expect short interarrival times and long service times to occur together more often than theory suggests. This would cause an upward bias in the waiting times and queue lengths observed in the simulation model.

In more complex models, the absence of independence among n-tuples is more difficult to assess. Verifying that a data source satisfies the independence assumption will always be of value, however, if an incorrect interpretation of results is to be avoided. References [5] and [16] offer helpful information to an experimenter in choosing a pseudorandom generator.

In some simulation experiments, correlated sampling is necessary. Suppose we are simulating the demand for aircraft tires; then tire wear-out is clearly related to the number of aircraft landings. Simulations of economic behavior often contain autocorrelated input processes, e.g. autonomous investment. References [7] and [21] describe methods for generating correlated samples and [19] describes procedures for sampling from two kinds of autocorrelated processes.

Tocher [24] has pointed out that correlated sampling is often difficult to perform because of the onerous and often impossible task of collecting sufficient information to describe desired distributions. Verification and validation should clearly be applied to correlated sampling. The peculiar circumstances surrounding different kinds of correlated sampling make it difficult to suggest a generally applicable method. Since all sampling ultimately depends on sequences of independent uniformly distributed random numbers, the least that can be done is to test the hypothesis that successive numbers and sequences of numbers are independent.

STRUCTURE VERIFICATION

Verifying the structure of simulation models means examining substructure outputs and determining whether they behave acceptably. One value of this exercise is that it identifies unwanted system behavior. Very minor simplifying assumptions can generate output processes whose behavior differs considerably from what is desired. Structure verification is also valuable for determining whether one may substitute an analytical or simple simulation substructure for a complex one. This may be done if a behavioral equivalence can be established between the simple and complex structures. The advantages of substitution accrue from the better understanding of the analytic or simple simulation structure and from savings in computation time during simulation.

To make behavioral comparisons, we require a probability model. The model must be sufficiently general to include the variety of phenomena encountered in simulation models, yet it must be restrictive enough to permit reasonably straightforward hypothesis testing. System simulations usually are concerned with series of interrelated events and an appropriate probability model must explicitly recognize interrelationships between past, present and future events. Since these associations are time-dependent, we refer to them as intertemporal dependence.

In Ref. [10], the writers suggest the class of covariance stationary stochastic processes as a convenient model for studying simulation-generated time series. The reasons for this choice are the valuable conceptual insights that these processes afford as well as the ease with which certain of their sample statistics (principally the spectrum) can be used in hypothesis testing. We first formally define a covariance stationary process and then discuss the meaning of some of its population parameters.

Let X_t be a random variable generated by a simulation model and recorded at time t. If $\{X_t;\ t = 0, \pm 1, \pm 2, \ldots, \pm \infty\}$ is a stochastic process such that $E(X_t X_{t+\tau})$ is finite and independent of t for all τ, then $\{X_t\}$ is <u>covariance stationary</u>. If the random variables X_t and $X_{t+\tau}$ are not independent for some $\tau \neq 0$, then $\{X_t\}$ is autocorrelated or linearly dependent. Output processes generally satisfy the covariance stationary assumptions and exhibit intertemporal dependence. The theory of covariance stationary processes provides a convenient framework within which to study the nature and extent of autocorrelation, the principal form of intertemporal dependence.

The <u>autocovariance function</u>

$$R_\tau = E(X_t X_{t+\tau}) - [E(X_t)]^2$$

summarizes all information concerning the autocorrelation present in $\{X_t\}$. The <u>spectrum</u>

$$g(\lambda) = (\pi)^{-1} \sum_{\tau=-\infty}^{\infty} R_\tau e^{-i\lambda\tau} \qquad 0 \leq \lambda \leq \pi$$

provides the same information, and in the writers' opinion is the preferred function to examine both for conceptual and statistical reasons [10].

The autocovariance function R measures the covariance between the random variables X_t and $X_{t+\tau}$. For the class of processes with which we are concerned this function diminishes, though not necessarily monotonically as $|\tau|$ increases. This property accords with reality, where the influence of the past wears off as time elapses. The spectrum g permits us to study mean-square variation in a series of interrelated events in terms of a continuum of frequency components. Since

$$R_0 = \int_0^{\pi} g(\lambda)d\lambda,$$

we may regard the variance R_0 as being made up of infinitesimal contributions $g(\lambda)d\lambda$ in small bands $d\lambda$ around each frequency. The spectrum g may be considered a variance decomposition with each component being associated with a specific frequency. Low frequencies correspond to long fluctuations in $\{X_t\}$; high frequencies correspond to rapid fluctuations. If a peak occurs in a spectrum, the corresponding frequency influences the appearance of $\{X_t\}$ to a greater extent than the remaining frequencies. A process with a peak at a non-zero frequency in fact displays something of a periodic appearance with its period corresponding approximately to the frequency at which the peak appears.

When the subscript t denotes time and X_t is an observation at time t, observations are collected at equal intervals on the time axis. Since t is only an index, it need not necessarily refer to time; any series of events can generate a time series. For example, in the simple queueing problem t may denote the t^{th} job to receive service and X_t may be the waiting time of this job. Here $\{X_t\}$ is a series of waiting times arranged in the order in which their corresponding jobs receive service.

Interactions between input and structure may often create unwanted periodicities in the output. This possibility is not as remote as one would like to think, for Slutsky [22] long ago showed that the linear summation of purely random events can appear regularly periodic. Since

peaks in a spectrum correspond to periodic components in $\{X_t\}$ and since the sharper a peak is, the more regular its periodicity is, examining the sample spectrum permits an experimenter to determine whether any periodicities exist and to estimate the extent of their regularity.

Figure 6 shows the sample spectrum of queue length for a single-server queueing model with exponentially distributed interarrival times and constant service time. The peak at 0.05 cycles per hour and its harmonics suggest the presence of periodicity. This behavior can be explained as follows. Whenever jobs are queueing, a periodic reduction in queue length occurs every 20 hours. With a constant service time, jobs emerge from the service facility at a fixed periodic rate, creating a periodic appearance. If this efflux is the input to another service facility, then this input is periodic whenever jobs are queueing in the first facility.

Two points motivate our concern about periodicities. First, their presence may be contrary to our intentions. Second, since the output of one substructure is usually the input to another, the effects of periodicity may propagate themselves throughout the remaining substructures. It is a property of substructures that they exhibit the characteristics of electromechanical systems and can have a natural or a resonant frequency. If a substructure is excited by a frequency close to its natural one, its response at that frequency is considerably exaggerated compared to that of others. The strength of a periodic component may therefore increase as it propagates through a system, obscuring the behavior of remaining components.

Conclusions drawn from the output of such a system may then be misleading. For example, one might suppose that the inputs to certain model subsystems are random phenomena whereas they actually appear in a model as regular or strongly periodic impulses. If this is so, rules appropriate for controlling randomly varying inputs may be judged inappropriate. The performance of the rules will be judged in an environment different from that for which they were designed.

As mentioned earlier, economy of detail aids understanding and saves computation time. The ease with which computer simulation languages permit one to describe complex behavior carries with it the

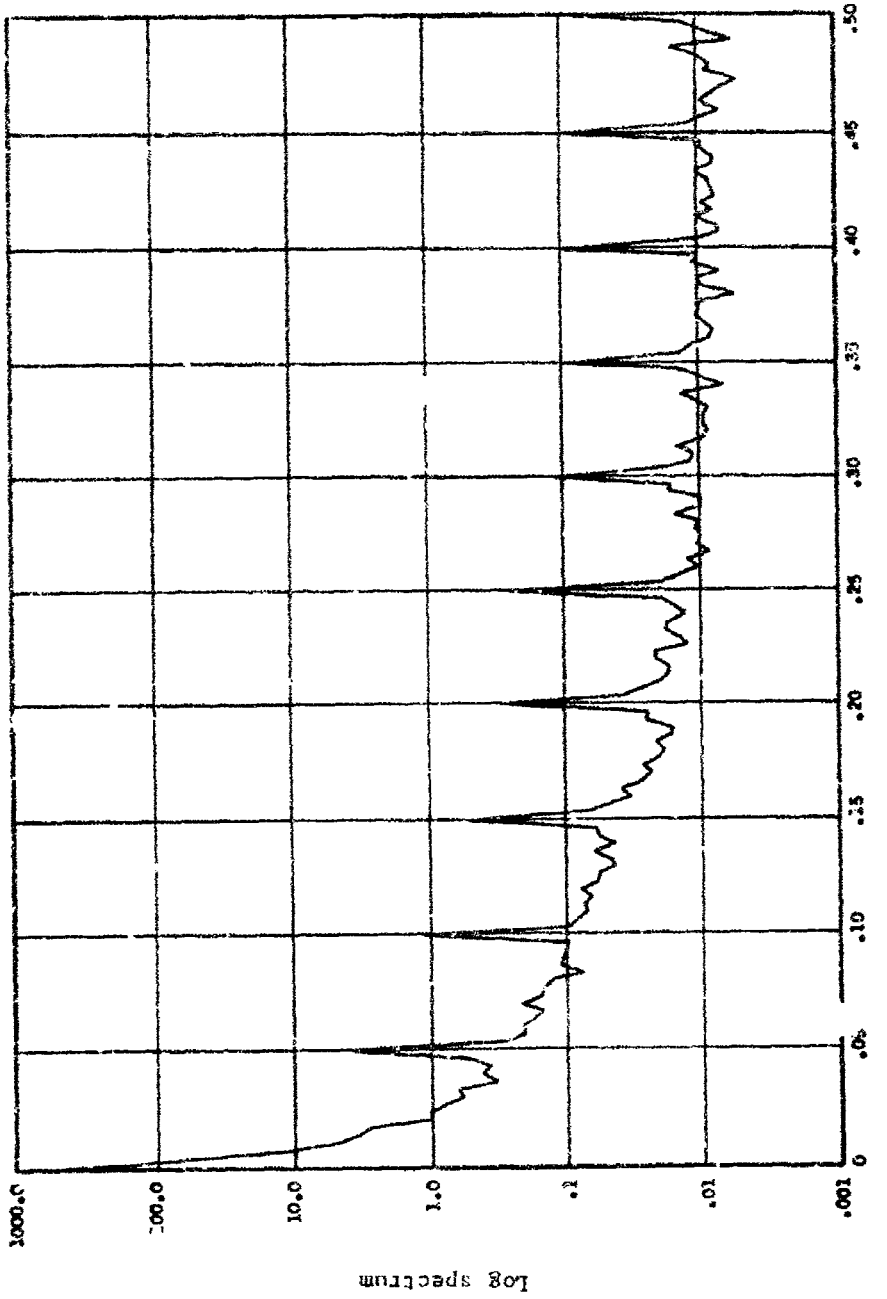

Fig. 6 -- Estimated queue-length spectrum with constant service time

danger of too much detail. Since a detailed model has more built-in assumptions than a simple model, it generally requires a longer learning period for a prospective user. In addition, simulating these details can consume vast amounts of computer time. If several models offer the same response to a given input, the simplest model is advantageous. It is desirable to test several models to determine the adequacy of each and then choose the simplest among the acceptable ones.

Suppose that a complex model behaves as required and we wish to test the equivalence of a simpler model. If at all possible, the simpler model should be compared with the true environment. When this cannot be done, the responses of the simple and complex models should be compared for a given input. The comparison tests the hypothesis that certain population characteristics, for example, means, variances or spectra, are identical for both models.

Since intertemporal dependence is often an important characteristic of models, and its mean-square variation is described by the spectrum, one may compare mean-square intertemporal dependence by testing the equivalence of spectra of two models. Jenkins [15] and Fishman and Kiviat [10] describe an appropriate testing procedure.

While it is true that higher-order effects may be dissimilar in the two models, a comparison of spectra can do much toward determining whether further comparisons are useful. The test is simple. In addition, when the null hypothesis of no difference is rejected, the comparison of spectra permits one to identify where in the structures of the two models the departures occur. With this knowledge, one may perhaps modify the simple structure to more closely match the complex one.

Verifying a model's structure protects an experimenter against creating anomalous responses, allows for a justifiably simple design, and saves computer time. It is a natural imperative to verify both data and structure before a model is used in order to minimize complications that can arise in the course of an experiment. Failure to verify has created more than one embarrassing situation in interpreting output.

IV. VALIDATION

DATA VALIDATION

Validating a model means establishing that it resembles its actual system reasonably well. If a model describes some hypothetical system, then no validation can occur. Also, if no numerical data exist for an actual system, it is not possible to establish the quantitative congruence of a model with reality. The ideas of this section therefore only apply when numerical data exist for some or all of an actual system.

Sampling from a theoretical rather than an empirical distribution is generally considered preferable, since it exposes a simulated system to the universe of possible stimuli rather than merely to those that have occurred in the past. Often, graphical methods suffice to judge the validity of theoretical distributions. If, for example, we assume that data have the exponential distribution, then we would expect the cumulative empirical distribution to appear linear on semilogarithmic paper. If the normal distribution is assumed, we would expect the cumulative empirical distribution to appear linear on normal probability paper. Graphic examination is easy and revealing. Whenever applicable, it should be used.

The χ^2 test is often proposed for testing the appropriateness of a chosen sampling distribution, but Cochran [2], among other writers, has shown the inadequacy of this test when the sample size of the empirical data is limited and the theoretical distribution is skewed. As an alternative, Cochran suggests the variance test, which generally has greater power than the χ^2 goodness-of-fit test and does away with the need for class intervals.

As an example, we describe the variance test when the null hypothesis is that a set of independent observations $\{x_i, i = 1, 2, \ldots, N\}$ came from an exponential distribution with parameter λ. Under this hypothesis we have

$$E(x_i) = 1/\lambda, \qquad \text{var}(x_i) = 1/\lambda^2.$$

As our estimate of λ we use the maximum likelihood estimator

$$\hat{\lambda} = N/\left(\sum_{i=1}^{N} x_i\right).$$

The test statistic is

$$\sum_{i=1}^{N} \frac{(x_i - 1/\hat{\lambda})^2}{(1/\hat{\lambda}^2)},$$

which is approximately distributed as χ^2 with (N-1) degrees of freedom. No class intervals are required in this test.

The χ^2 and variance tests both assume independent observations, an assumption that also simplifies Monte Carlo sampling. While its convenience for testing is apparent, the credibility of this assumption is seldom tested. If a sample record is "sufficiently long," one may estimate its spectrum and compare it with the uniform spectrum for an uncorrelated process.

For short records, spectrum comparisons are not possible. Here we suggest using nonparametric tests of randomness which do not require an investigator to make any assumptions about the underlying distribution of sample data. In addition, the appropriateness of the tests do not depend on the sample being large. Walsh [25] lists a number of nonparametric tests that can be applied to small samples.

The term "sufficiently long" has an irritating quality about it for simulation experimenters. Seldom is enough prior information available to estimate how long to run an experiment. Nevertheless, most writers on the statistical analysis of simulation experiments take the length of the sample record as adequate for the analyses they propose. In [9] a two-stage technique is described wherein one may estimate how long an experiment is to be run. The procedure is integrated into a test comparing means, but this should pose no problem in determining run lengths alone.

STRUCTURE VALIDATION

Having tested assumptions about the data, there remains the task of validating the structure. If a model resembles reality fairly well, we expect that its simulated response to a simulated, but valid, input should exhibit behavior similar to that observed for the real system. A spectrum analysis is again instructive. Testing the homogeneity of spectra, one for the actual system's output and the other for the simulated system's output, is easily accomplished as described in [10] and [15].

The spectrum comparison applies to testing the homogeneity of the autocorrelation structure. Comparing means is also desirable since we would expect no difference if the simulation model adequately resembles the true system. Since the output processes are generally autocorrelated, a comparison of means requires more work and care than in the case of independent observations.

The procedures in [8] can easily be modified to compare the means of the simulated and real systems. The variance of the sample mean is shown to be proportional to the spectrum at zero frequency and, hence, testing means and testing spectra show a number of common features.

Validation, while desirable, is not always possible. Each investigator has the soul-searching responsibility of deciding how much importance to attach to his results. When no experience is available for comparison, an investigator is well advised to proceed in steps, first implementing results based on simple well-understood models and then using the results of this implementation to design more sophisticated models that yield stronger results. It is only through gradual development that a simulation can make any claim to approximate reality. Large scale models that are not amenable to validation often lead to perplexing, if not misleading, results. This occurs partly because the complexity of a system confuses a model-builder and partly because of the tenuous nature of results based on cascaded approximations. Despite its difficulty, effort must be expended on model validation -- first, to give credence to results

within the validated range of model operations, and second, to instill confidence in extrapolations beyond the range of model experience.

Verifying and validating a model comprise but a small share of the statistical problems in a simulation experiment. Once an experimenter accomplishes them, he can begin to exercise his model to get answers. His purpose is to collect data, reduce them, and make inferences about them, as efficiently as possible. We classify the statistical problems he encounters under problem analysis. The way he solves these problems strongly influences the quality of his results.

V. PROBLEM ANALYSIS

One purpose of system simulation experiments is to compare system responses to different operating rules. In the simple queueing problem, for example, we may wish to compare the mean queue lengths caused by given arrival and service rates when different rules are used to assign priorities to jobs. Another purpose is to determine functional relationships between input factors and system response. We may simply wish to get a "feel" for the way in which input and output relate, or we may wish to use a determined functional relationship in a further analysis. For example, we may determine a functional relationship when all inputs are unrestricted and then use this relationship to find the maximum response when constraints are placed on the inputs. In some studies, both purposes enter. For simplicity, we treat them separately.

Regardless of purpose, there are several statistical questions common to all problem analyses and to structural verification and validation as well. One question relates to the choice of sampling interval: What is the proper interval of simulated time between successive observations of a process of interest? Another question is: How can results be obtained efficiently with a given reliability? This topic is often discussed under the heading of "variance reduction techniques." Reliability estimation itself poses another statistical problem in system simulation experiments that must be solved before one can determine how long to run an experiment.

Other statistical questions are peculiar to particular kinds of experiments. When comparing experiments, one requires statistical testing procedures. When relating response to input, one asks where in the input ranges it is best to measure response so that its functional form can be most easily identified and its parameters most reliably estimated.

Measurements made in a simulation experiment can be of two kinds. One kind measures a system's response to all possible situations. Here the relevant statistic is a time-integrated average. The other measures a system's response to a specific set of initial

conditions. Time-integrated averages appear to be the most common measurement. The simulation literature is principally concerned with them, and the discussion here retains this emphasis. The reader should not conclude from this that measurement of response to initial conditions is unimportant. In particular the lack of literature on the subject should be taken as a comment on its specialized nature, not its worth. As the use of simulation increases, there will be more concern for measurements of this kind and more will be written about them. As indicated, the discussion from here on will be of experiments performed with the first kind of measurements in mind. The remainder of this section uses the terms "time-integrated average" and "sample mean" interchangeably.

SAMPLING INTERVAL

When t denotes time the meaning of a time-integrated average is clear. When t is a more general ordering index, a time-integrated average refers to the mean value of a quantitative characteristic of a series of events indexed on t. The term time-integrated average remains appropriate since the ordering of events is related to time.

If the index denotes time then the choice of sampling interval is crucial if we hope to extract useful information about the autocorrelation structure of a process in an efficient way. For our purposes a sampling interval should be small enough so that within it a process changes little, if at all. Process activity, not chronological time, dictates the choice of sampling interval. For each experiment, other than replications, it is wise to check the adequacy of the sampling interval, since too small an interval causes redundancy in the data and too large an interval loses information. Biasing an interval downward is more desirable than biasing it upward, since redundant data are far less harmful than lost information.

When t denotes an event in an ordered series, the role of the sampling interval is changed. Since we simply collect an observation every time an event occurs, it would seem that we could avoid choosing a sampling interval. It may occur, however, that successive events are so highly correlated that collecting information on each event is highly

redundant. When this is the case a judicious choice of sampling interval reduces the number of observations without sacrificing any significant information.

VARIANCE REDUCTION TECHNIQUES

It is naturally of interest to obtain experimental results with specified reliability at minimum cost. Development of variance reduction techniques was in fact the principal statistical activity in the early days of computer simulation (Monte Carlo) experiments. The importance of this activity continues to grow with the increasing complexity of experiments and their concomitant consumption of computer time.

Hammersley and Handscomb [11] discuss several variance reduction techniques, among which the method of antithetic variates appears easiest to apply. Page [20] shows its use in a simulated queueing problem. Briefly, by generating ξ, a uniformly distributed random number, in one replication of an experiment and generating $1-\xi$ in a second replication, the method induces negative correlation between the responses obtained in both replications. The variance of the average response of the two replications is consequently smaller than it would be if the replications were independent. Antithetic variates may also be used with more than two replications.

When the comparison of experiments is the purpose of a simulation exercise, one may improve the efficiency of the data-gathering procedure in another way. When testing the difference of two means, for example, one may reduce the variance of the difference by choosing the sample sizes as functions of the variances of the individual sample means, the computer times required to collect one observation in each experiment, and the degree of correlation between the sample means. Inducing a positive correlation between the sample means reduces the variance of their difference. This can be done, in some cases, by using the same set of random numbers for both experiments.

As mentioned in Sec. IV, the choice of the number of observations to collect in each experiment is a major influence in minimizing the computer time needed to meet a specified level of accuracy. The

two-stage procedure given in [9] offers a reasonably straightforward way of coming close to the most efficient sample sizes. When the sample sizes are chosen close to the efficient solution, a major saving in computer time accrues.

ESTIMATING RELIABILITY

Since experimental results are random variables, it is important that their reliability as estimates of population parameters be stated explicitly. Failure to do so obscures the fact that some results may be better than others. In addition, omitting reliability measures makes it impossible to determine how much longer to run an experiment in order to improve its reliability by some fixed amount.

Variance reduction techniques permit us to reduce the computer time necessary to obtain a result with a given reliability. We must also have a way of estimating the reliability of a result. This has long been a major problem area in simulation experiments.

If a sampling interval is chosen so that observations are independent, then the variance of a time-integrated average or sample mean is simply the population variance divided by T, the number of observations. In general, since simulation data are autocorrelated, the above approach requires finding a sampling interval such that successive observations are reasonably independent. Mechanic and McKay [18] have investigated this approach.

If, however, one treats a simulated process as a covariance stationary stochastic process (which it generally is), then the variance of the sample mean is $\pi g(0)/T$ where the function g is defined in Sec. III, and T is the length of the simulation run. A procedure for estimating g(0) is given in [8], but unfortunately it cannot easily be incorporated into the experimental run itself.

Another approach is to sum sample means from independent replications of the same experiment. The variance of this sum is, of course, inversely proportional to the number of replications. Using antithetic variates can reduce the variance even more by inducing negative correlation between sample means.

COMPARISON OF EXPERIMENTS

In an experiment, response is generally a large-sample, time-integrated average that satisfies the conditions for asymptotic normality. This fact greatly simplifies testing the difference of two means obtained under different operating rules, since the difference of the sample means is also asymptotically normal. Let the subscripts 1 and 2 denote experiments 1 and 2, respectively. Then for a given significance level α and tolerance δ, we have, under the null hypothesis of no difference in the means,

$$\text{prob}\ (|\bar{X}_1 - \bar{X}_2| \leq \delta) = 1 - \alpha.$$

To test the null hypothesis, we require reasonably accurate estimates of the variances of the sample means. These can be obtained by procedures described in [9].

The comparison just described is the one most commonly applied in the analysis of experimental results. Multiple comparisons and ordering procedures are desirable when more than two sets of operating rules are being considered. Their appropriate statistical procedures are found in texts on the analysis of variance. To our knowledge, no study has yet appeared that makes a substantive contribution toward adapting these procedures to the peculiar environment of computer simulation experiments.

RESPONSE MEASUREMENT

In comparing experiments, one is concerned with the response of a system to different qualitative factors, such as operating rules. Alternatively, one may examine the system's response under given operating rules to changes in quantitative factors, such as different input activity levels. We refer to this analysis as response measurement. Its purpose is to find a functional form relating the variable parameters of an input process to an observed output, and to estimate the coefficients of the functional form.

Consider a simulation with one input x and one output y. For each experiment, x assumes a fixed value that is known exactly, whereas y assumes a value from a probability distribution whose parameters are functions of x; y is a random variable. In a queueing problem x might be the mean arrival rate and y the sample mean number of jobs in queue. If, for estimation purposes, we use the linear least-squares method, our functional relationship for the i^{th} observation is

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

$$X_i = f(x_i)$$

$$Y_i = f(y_i).$$

To derive the best linear unbiased estimates of α and β with the linear least-squares method, we require

$$E(\epsilon_i) = 0$$

$$E(\epsilon_i \epsilon_j) = 0, \quad i \neq j$$

$$\text{var}(Y_i) = c \quad \text{for all } i.$$

Some commonly used functional forms are listed below.

	f(x)	f(y)
1	x	y
2	log x	y
3	x	log y
4	log x	log y

For a correctly chosen form, the relationship between f(x) and f(y) will appear linear. Linear, semilog, and log-log graph paper may be used to find which relationship is most appropriate. Other forms may be examined, but for the moment we assume that one of the above forms will hold. Hoerl [14] describes several techniques for identifying the functional form that linearizes the relationship between x and y.

It is convenient to distinguish between two kinds of observations: those collected to determine the appropriate functional form, and those collected to estimate α and β with a given level of accuracy. The first set is a subset of the second.

To satisfy the above regression model, we require all y_i's to be independent and have a common variance. Independence can be gained by using different random number sequences in successive simulation runs with each set of input activity levels. For a given variance, the proper length of a simulated run may be estimated by the two-stage procedure to which we have already alluded.

To find a functional form it is necessary to take observations for a number of input activity levels within the range of activity levels being considered. As one would expect, the number of such observations is inversely related to the variance of the observations. The more reliable the observations, the more confidence one can place in having identified an appropriate functional form for a given number of observations.

Once an appropriate functional form is found, one uses the observations already collected to estimate the coefficients. Additional observations may be collected and used to improve the reliability of the estimates . The objective at this step is efficiency -- the conservation of computer time. If the computer times required to collect all y_i's with equal variance are the same, taking additional observations at the ends of the x range minimizes the computer time necessary to improve the reliability of the estimated coefficients by a given amount. In general, the computer times required to collect observations with common variance do differ and, hence, the choice of where to collect observations is not so simple. Little, if anything, has been published about this problem. Its solution will undoubtedly improve the efficient performance of simulation experiments.

It may occur that a priori theory suggests a model of the form,

$$Y_i = \alpha + \sum_{j=1}^{N} \beta_j X_i^j + \epsilon_i.$$

This model, unlike the one above, does not exclusively take observations at the end points of the independent variable range to minimize the sample size needed for a given accuracy. The points at which observations should be taken are given by the zeros of a polynomial which is the integral of one of the Legendre polynomials [13].

Response surface exploration, optimum seeking methods, and sequential experimentation are all topics germane to the analysis of computer simulation experiments. Cochran and Cox [3] describe the principles of response surface methodology, and Hill and Hunter [12] list a number of papers covering different aspects of the topic. Draper and Smith [6] describe procedures for applying a variety of linear regression analyses. Wilde [26] describes simple methods for finding maxima and minima. Cochran and Cox also discuss sequential experimentation. Although these methods contribute significantly to the statistical analysis of experiments, they remain to be integrated into a general procedure that takes due cognizance of the peculiarities of computer simulation experiments.

REFERENCES

1. Cochran, W. G., "The χ^2 Test of Goodness of Fit," Ann. Math. Stat., Vol. 23, No. 3, September 1952, pp. 315-345.

2. -----, "Some Methods for Strengthening the Common χ^2 Test," Biometrics, Vol. 10, No. 4, December 1954, pp. 417-451.

3. -----, and G. M. Cox, Experimental Designs, John Wiley and Sons, New York, 1957.

4. Colker, A., et al., The Generation of Random Samples from Common Statistical Distributions, United States Steel Corporation, Applied Research Laboratory Report 25.17-016(1), November 1962.

5. Coveyou, R. R., and R. D. MacPherson, "Fourier Analysis of Uniform Random Number Generators," J. Assoc. Comp. Mach., Vol. 14, No. 1, January 1967, pp. 100-119.

6. Draper, N. R., and H. Smith, Applied Regression Analysis, John Wiley and Sons, New York, 1966.

7. Fieller, E. C., T. Lewis, and E. S. Pearson, Correlated Random Normal Deviates, Tracts for Computers No. 26, Cambridge University Press, London, 1955.

8. Fishman, G. S., Problems in the Statistical Analysis of Simulation Experiments: The Comparison of Means and the Length of Sample Records, The RAND Corporation, RM-4880-PR, February 1966. Also published in Comm. ACM, Vol. 10, No. 2, February 1967, pp. 94-99.

9. -----, Digital Computer Simulation: The Allocation of Computer Time in Comparing Simulation Experiments, The RAND Corporation, RM-5288-1-PR, October 1967. Also to be published in J. Oper. Res.

10. -----, and P. J. Kiviat, "The Analysis of Simulation Generated Time Series," Mgt. Sci., Vol. 13, No. 7, March 1967, pp. 525-557.

11. Hammersley, J. M., and D. C. Handscomb, Monte Carlo Methods, Methuen, London, 1964.

12. Hill, W. J., and W. G. Hunter, "A Review of Response Surface Methodology: A Literature Survey," Technometrics, Vol. 8, No. 4, November 1966, pp. 571-590.

13. Hoel, P., "Efficiency Problems in Polynomial Regression," Ann. Math. Stat., Vol. 29, No. 4, December 1958, pp. 1134-1145.

14. Hoerl, A. E., Jr., "Fitting Curves to Data," in J. W. Perry (ed.), Chemical Business Handbook, McGraw-Hill Book Company, Inc., New York, 1954.

15. Jenkins, G. M., "General Considerations in the Analysis of Spectra," Technometrics, Vol. 3, No. 2, May 1961, pp. 133-166.

16. MacLaren, M. D., and G. Marsaglia, "Uniform Random Number Generators," J. Assoc. Comp. Mach., Vol. 12, No. 1, January 1965, pp. 83-89.

17. Mann, H. B., and A. Wald, "On the Choice of the Number of Class Intervals in the Application of the Chi Square Test," Ann. Math. Stat., Vol. 13, 1942, pp. 306-317.

18. Mechanic, H., and W. McKay, Confidence Intervals for Averages of Dependent Data in Simulations II, IBM Advanced Systems Development Division, No. 17-202, Yorktown Heights, New York, 1966.

19. Naylor, T. H., et al., Computer Simulation Techniques, John Wiley and Sons, New York, 1966

20. Page, E. S., "On Monte Carlo Methods in Congestion Problems, II: Simulation of Queueing Systems," J. Oper. Res., Vol. 13, No. 2, March-April 1965, pp. 300-305.

21. Rakov, G. K., Generation of a Random Correlated Quantity on a High-Speed Electronic Computer, translated from Russian, U.S. Joint Publ. Res. Service, JPRS:5784, November 1960.

22. Slutzky, E., "The Summation of Random Causes as the Source of Cyclic Processes," Econometrica, Vol. 5, 1937, pp. 105-146.

23. Symposium on Monte Carlo Methods, John Wiley and Sons, New York, 1955.

24. Tocher, K. D., The Art of Simulation, Van Nostrand, Princeton, New Jersey, 1963.

25. Walsh, J., Handbook of Nonparametric Statistics, Van Nostrand, Princeton, New Jersey, 1962.

26. Wilde, D. J., Optimum Seeking Methods, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

DOCUMENT CONTROL DATA

1. ORIGINATING ACTIVITY	2a. REPORT SECURITY CLASSIFICATION
THE RAND CORPORATION	UNCLASSIFIED
	2b. GROUP

3. REPORT TITLE

DIGITAL COMPUTER SIMULATION: STATISTICAL CONSIDERATIONS

4. AUTHOR(S) (Last name, first name, initial)

Fishman, George S. and Philip J. Kiviat

5. REPORT DATE	6a. TOTAL No. OF PAGES	6b. No. OF REFS.
November 1967	41	26

7. CONTRACT OR GRANT No.	8. ORIGINATOR'S REPORT No.
F44620-67-C-0045	RM-5387-PR

9a. AVAILABILITY/LIMITATION NOTICES	9b. SPONSORING AGENCY
DDC-1	United States Air Force Project RAND

10. ABSTRACT

A discussion of the statistical problems that arise in computer simulation experiments. Three problem areas inherent in all stochastic system simulation models are discussed: verification, which determines whether a model actually behaves as an experimenter assumes it does; validation, which tests whether the model reasonably approximates a real system; and problem analysis, which seeks to ensure proper execution of a simulation and proper handling of its results. The study traces the elements of a simulation experiment from initial conception to analysis of final results, defining the statistical problems that arise at each step and relating them to the formal body of statistical theory. Since the aim is to promote awareness of problems, not to solve them, the study offers no general solutions but provides references germane to the statistical problems described.

11. KEY WORDS

Computer simulation
Statistical methods and processes